Article

Not peer-reviewed version

Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review

Fnu Neha * , Deepshikha Bhati , Deepak Kumar Shukla

Posted Date: 13 August 2025

doi: 10.20944/preprints202508.1022.v1

Keywords: retrieval-augmented generation (RAG); large language models (LLMs); biomedical natural language processing (NLP); artificial intelligence (AI); medical question answering; domain-specific language models; healthcare AI

Article

Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review

Fnu Neha [1,*], Deepshikha Bhati [1] and Deepak Kumar Shukla [2]

[1] Department of Computer Science, Kent State University, Kent, OH 44242
[2] Rutgers Business School, Rutgers University, Newark, NJ 07102
[*] Correspondence: neha@kent.edu

Abstract

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by integrating external knowledge retrieval to improve factual consistency and reduce hallucinations. Despite growing interest, its use in healthcare remains fragmented. This review fills that gap through an analysis of 30 peer-reviewed studies on RAG in clinical domains, focusing on three of its most prevalent and promising applications in diagnostic support, electronic health record (EHR) summarization, and medical question answering. We synthesize existing architectural variants (naïve, advanced, and modular) and examine their deployment across these applications. Persistent challenges are identified, including retrieval noise (irrelevant or low-quality retrieved information), domain shift (performance degradation when models are applied to data distributions different from their training set), generation latency, and limited explainability. Evaluation strategies are compared using both standard metrics and clinical-specific metrics, FactScore, RadGraph-F1, and MED-F1, which are particularly critical for ensuring factual accuracy, medical validity, and clinical relevance. This synthesis offers a domain-focused perspective to guide researchers, healthcare providers, and policymakers in developing reliable, interpretable, and clinically aligned AI systems, laying the groundwork for future innovation in RAG-based healthcare solutions.

Keywords: retrieval-augmented generation (RAG); large language models (LLMs); biomedical natural language processing (NLP); artificial intelligence (AI); medical question answering; domain-specific language models; healthcare AI

1. Introduction

Artificial intelligence (AI) plays an increasingly important role in healthcare, supporting tasks such as diagnosis, treatment planning, patient safety, risk prediction, medical imaging, triage, and clinical documentation. However, as electronic health records (EHRs) continue to grow in scale and clinical knowledge expands rapidly, accessing accurate and up-to-date information at the point of care remains a significant challenge.

Recent advances in large language models (LLMs), such as OpenAI's GPT-4.0, Claude 3, Large Language Model Meta AI (LLaMA) 3, Qwen 2.5, and Gemini 2.0 [1], show improved reasoning, reduced latency, and multimodal capabilities. While these general-purpose models perform well on standard benchmarks, hallucinations and misalignment with clinical knowledge remain significant concerns when deployed without domain-specific grounding.

To address these challenges, medical domain-specific LLMs such as Medical Pathways Language Model (Med-PaLM) [2] and Biomedical Generative Pretrained Transformer (BioGPT) [3] have shown improved performance across various clinical natural language processing (NLP) tasks, including question answering, summarization, entity recognition, relation extraction, document classification, and literature mining. However, despite their domain specialization, these models are trained on static datasets and lack access to real-time or external knowledge sources, which can lead to hallucinations and outdated responses [4]. This limitation raises concerns about factual accuracy, reliability, and traceability in high-stakes clinical applications.

Many LLMs have been fine-tuned on biomedical corpora to enhance domain-specific comprehension. Variants of Bidirectional Encoder Representations from Transformers (BERT) [5], such as Biomedical BERT (BioBERT) [6] and ClinicalBERT [7], improve contextual understanding of medical terminology. However, these models cannot incorporate continuously evolving clinical knowledge, limiting their adaptability in dynamic healthcare environments.

Introduced in 2020, Retrieval-Augmented Generation (RAG) mitigates these limitations by combining a retriever which accesses external sources such as clinical guidelines or EHRs with a generator that conditions responses on both the query and retrieved content [8]. This architecture enables integration of up-to-date information, enhances transparency, reduces hallucinations, and produces evidence-grounded outputs, making RAG well-suited for medical applications.

Despite its advantages, deploying RAG in clinical settings presents challenges, including retrieval noise (e.g., surfacing generic content instead of condition-specific guidelines), domain shift (e.g., mismatched terminology between EHRs and biomedical literature), inference latency (e.g., delays in generating responses during emergencies), and limited interpretability (e.g., unclear how retrieved text shapes treatment suggestions) [9]. Additionally, ethical concerns such as privacy, data security, bias, and regulatory compliance are also needed to be addressed, alongside continuous validation, governance, and feedback mechanisms for trustworthy deployment.

While interest in RAG continues to grow, most prior work has focused either on architectural innovations [10] or on isolated components such as retriever and generator design [11]. A comprehensive, domain-specific review that integrates both architectural methodologies and clinical implementations remains lacking. Thus, this paper fills this gap by providing a structured review of RAG techniques applied in the medical domain. It examines architectural variants, evaluation strategies, practical deployments, and integration challenges, highlighting their implications for AI-driven healthcare. The paper concludes with a synthesis of open problems and future directions for safe, scalable, and effective RAG in the clinical domain. The key contributions of this paper are:

1. An overview of the RAG architecture and its advantages over standalone language models.
2. A survey of RAG applications in healthcare across tasks such as question answering, summarization, and evidence retrieval.
3. A review of domain specific and standard evaluation metrics.
4. A detailed discussion of challenges, including retrieval instability, generation latency, domain shift, and limited transparency.
5. A synthesis of emerging directions such as multimodal retrieval, continual learning, federated architectures, and clinically aligned evaluation strategies.

The remainder of this paper is structured as follows: Section 2 outlines the research methodology. Section 3 provides background on NLP, LLMs, and RAG. Section 4 explores related work. Section 5 discusses RAG-based healthcare applications. Section 6 discusses evaluation metrics and benchmarks. Section 7 addresses challenges and limitations. Section 8 discusses future research directions, and Section 9 concludes the paper.

2. Research Methodology

This review follows a structured methodology to examine the use of RAG in healthcare. It includes searches across databases, application of inclusion and exclusion criteria, article screening, and taxonomy-based classification. A comprehensive search was performed in PubMed, IEEE Xplore, Scopus, Web of Science, Google Scholar, and ACM Digital Library, covering publications from 2022 to 2025. The search query combined the terms: `("Retrieval-Augmented Generation" OR "RAG") AND ("healthcare" OR "clinical applications") AND ("language models" OR "LLMs") AND ("question answering" OR "summarization" OR "knowledge grounding")`.

2.1. *Inclusion and Exclusion Criteria*

Inclusion:

- Studies applying RAG or retrieval-augmented LLMs within healthcare domains.
- Peer-reviewed journal articles, conference proceedings, or high-quality preprints published in English.
- Publications presenting empirical results or detailed implementation frameworks.

Exclusion:

- Studies unrelated to healthcare or not utilizing RAG-based models.
- Editorials, opinion articles, or conceptual papers lacking experimental validation.
- Publications without full-text access or written in languages other than English.

2.2. *Data Retrieval and Screening*

The initial search yielded 123 records. After removing 46 duplicates, 77 were screened by title and abstract; 47 were excluded. The remaining 30 underwent full-text review. All 30 were included, as two reviewers independently screened full texts, with conflicts resolved by consensus. Figure 1 shows the PRISMA-based selection flow.

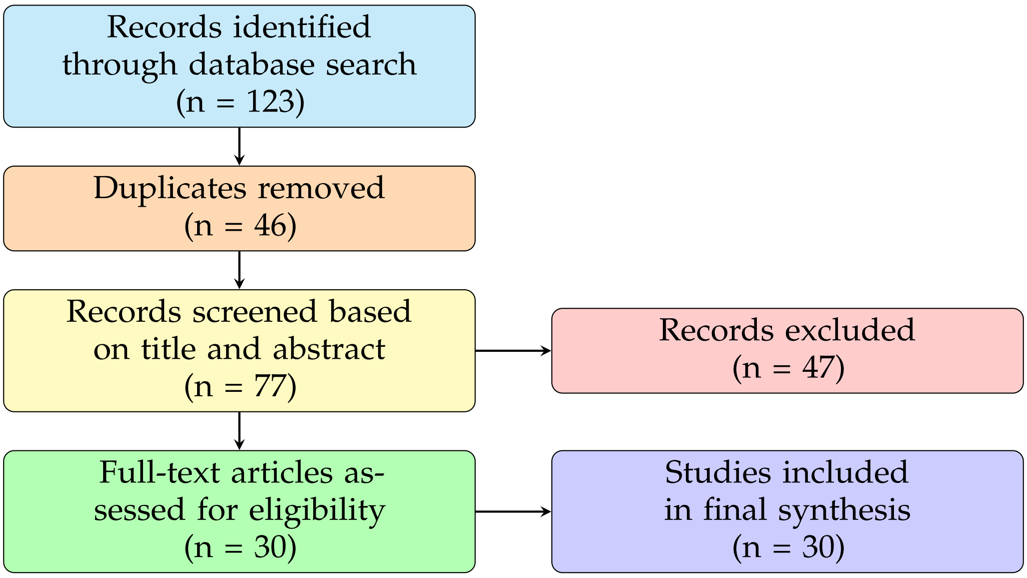
Records identified through database search (n = 123)
Duplicates removed (n = 46)
Records screened based on title and abstract (n = 77)
Records excluded (n = 47)
Full-text articles assessed for eligibility (n = 30)
Studies included in final synthesis (n = 30)

Figure 1. PRISMA-style flow diagram of study selection process.

2.3. *Taxonomy Development*

A taxonomy was constructed to categorize the selected studies by their primary application domains, as shown in Table 1.

2.4. *Research Objectives (RO)*

This review is guided by the following objectives:

- **RO1:** To assess the effectiveness of RAG in supporting clinical workflows and enhancing interpretability.
- **RO2:** To identify key gaps in current literature and propose directions for future RAG research that align with clinical needs, safety, and explainability.

Table 1. Taxonomy of RAG Applications in Healthcare.

No.	Category	Description
1	General Clinical Applications of RAG	Broad applications of RAG for tasks like clinical summarization, decision support, and guidelines.
2	RAG Chatbots for Patient Interaction	Conversational agents enhanced by retrieval for providing personalized medical advice.
3	Specialty-Focused RAG Models	RAG frameworks tailored for domains such as cardiology, nephrology, or oncology using specialty-specific knowledge bases.
4	RAG for Signal and Time-Series Tasks	Integration of RAG with biosignals like ECG, EEG, or wearable data for diagnostic interpretation.
5	Graph-Based and Ontology-Aware RAG Frameworks	Use of structured clinical ontologies or knowledge graphs for enhanced retrieval and explainability.
6	RAG with Blockchain and Secure Architectures	Incorporation of privacy-preserving, decentralized data retrieval using blockchain-enhanced architectures.
7	Radiology-Specific Retrieval-Augmented QA	RAG systems designed for image-report alignment, report generation, and visual question answering in radiology.

3. Background and Fundamentals

Early NLP systems in healthcare were rule-based, relying on symbolic logic, pattern matching, and expert-defined ontologies for tasks such as symptom triage, medical coding, and information extraction [12]. Although effective in narrow, controlled environments, these systems struggled with ambiguous or context-dependent language and were highly sensitive to variations in clinical documentation. Moreover, these lacked scalability across institutions and medical specialties.

The advent of machine learning (ML) introduced statistical models that improved adaptability and performance. However, these models required structured inputs and extensive manual feature engineering to extract meaningful signals from clinical text [13]. Deep learning further advanced the field by enabling end-to-end processing of unstructured data, with early implementations using recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks [14]. Despite these improvements, these architectures struggled with long-range dependencies and parallel computation.

Transformers, introduced by Vaswani et al. in the 2017 paper "Attention is All You Need" [15], addressed the limitations of earlier architectures by introducing a self-attention (SA) mechanism capable of modeling both local and global dependencies within a sequence. SA enables the model to weigh the relevance of each token relative to others, allowing it to focus on contextually important information regardless of position. To enhance representational capacity, Transformers employ multi-head attention, which runs multiple attention operations in parallel. Each head captures different relationships or dependencies, and the outputs are concatenated and linearly transformed. This parallel structure improves scalability, enables efficient parallel computation, and strengthens the model's ability to capture long-range context.

LLMs are built on transformer architectures and pretrained using self-supervised objectives such as masked language modeling (MLM) or causal language modeling (CLM) [16]. MLM (used in BERT-like models) involves masking random tokens in the input and training the model to predict them based on the surrounding context. In contrast, CLM (used in GPT-like models) trains the model to predict the next token in a sequence, relying on prior context in an autoregressive manner.

When pretrained on large-scale biomedical corpora such as the medical information mart for intensive care (MIMIC) -III clinical data [17] or medical literature, LLMs acquire domain-specific language understanding. Biomedical models like BioGPT, optimized for literature-based question answering, and ClinicalT5 [18], pretrained on discharge summaries, have shown strong performance on tasks including named entity recognition, question answering, document classification, and summarization.

Despite their capabilities, LLMs encode knowledge implicitly within their parameters, making it difficult to update or verify information post-training.

3.1. Retrieval-Augmented Generation: Foundations

Retrieval-Augmented Generation (RAG) mitigates key limitations of standalone LLMs by integrating non-parametric external memory through a retriever module for enhancing factual grounding. It improves knowledge coverage and enables dynamic integration of evolving clinical content. A RAG system consists of two main components: a retriever R that identifies relevant documents from a corpus $\mathcal{D}$, and a generator G that conditions on both the input query and the retrieved context to produce the output.

Given a query x, the retriever selects the top-k documents $\{d_1, \ldots, d_k\} \subset \mathcal{D}$ using a similarity function, defined as:

$$\mathrm{sim}(x, d_i) = f_q(x)^\top f_d(d_i) \tag{1}$$

where f_q and f_d are neural encoders for the query and documents, respectively. Similarity is computed using the dot product or cosine similarity in a shared embedding space.

The generator produces output y by estimating the conditional probability over possible responses.

$$P(y \mid x) = \sum_{i=1}^{k} P(d_i \mid x) \cdot P(y \mid x, d_i) \tag{2}$$

Here, $P(d_i \mid x)$ can be derived from normalized similarity scores or treated uniformly in top-k retrieval. The final output can be generated either by marginalizing across retrieved documents or by concatenating them as a single input:

$$y = \arg\max_{y'} P(y' \mid x, \{d_1, \ldots, d_k\}) \tag{3}$$

RAG retriever operates in one of two modes:

- **Sparse Retrieval:** This approach relies on lexical overlap between the query and documents. Common methods include Term Frequency-Inverse Document Frequency (TF-IDF) [19] and BM25 [20]. TF-IDF assigns higher importance to terms that appear frequently in a document but rarely across the corpus, while BM25 improves upon TF-IDF by incorporating term frequency saturation and document length normalization. Sparse retrievers are fast, interpretable, and require no training. However, they struggle to capture semantic similarity and are sensitive to lexical variations. It is an important limitation in the medical domain, where abbreviations, synonyms, and varied terminology are frequent.
- **Dense Retrieval:** Unlike sparse retrieval, dense retrievers use neural encoders to map both queries and documents into a shared embedding space, enabling semantic similarity matching. These models are typically trained on large datasets using contrastive learning objectives, allowing them to capture meaning beyond exact word overlap. Popular dense retrievers include dense passage retrieval (DPR) [21]. In clinical settings, dense retrieval is useful for handling synonyms, abbreviations, and contextually rich queries. However, dense retrievers are more computationally expensive, require training data, and can be less interpretable than their sparse counterparts.

Table 2. Comparison of Retrieval Strategies.

Key Points	TF-IDF	BM25	DPR
Type	Sparse (lexical)	Sparse (lexical)	Dense (neural)
Retrieval Method	Term frequency weighting	Probabilistic scoring with term normalization	Bi-encoder with semantic embeddings
Similarity Metric	Cosine similarity	BM25 score	Dot product or cosine similarity
Training Requirement	None	None	Supervised (Q-A pairs)
Context Sensitivity	Low	Moderate	High
Efficiency	Fast	Fast	Moderate (GPU preferred)
Scalability	High	High	Moderate
Memory Usage	Low	Low	High
Output Quality	Lexical match only	Improved over TF-IDF via ranking	Context-aware semantic relevance
Typical Use Cases	Baseline IR, filtering	Search engines, ranking tasks	RAG, QA, chatbots
Dependencies	Bag-of-words model	Bag-of-words + heuristics	Pretrained LLMs (e.g., BERT)

Generator models such as Text-to-Text Transfer Transformer (T5) [22], GPT, and BioGPT are adapted to receive both the input x and retrieved documents d_i through input concatenation or cross-attention mechanisms within the transformer decoder. These models are trained using maximum likelihood estimation (MLE) to predict the next token given the input and context, or via reinforcement learning with human feedback (RLHF) to optimize for attributes such as factuality, readability, or clinical alignment. RAG frameworks enable grounded generation by aligning internal parametric knowledge with external, dynamically retrieved evidence.

Table 3. Comparison of Sparse vs. Dense Retrieval.

Key Points	Sparse Retrieval	Dense Retrieval
Retrieval Mechanism	Lexical token overlap	Learned embedding similarity
Input Representation	Bag-of-Words (BoW) vectors	Neural embeddings (contextual)
Similarity Metric	BM25, TF-IDF (exact match)	Dot product or cosine similarity
Training Requirement	No training needed	Requires supervised training
Speed	Fast (index lookup)	Slower (approximate nearest neighbor search)
Semantic Matching	Low (sensitive to term variation)	High (captures semantic context)
Memory Usage	Low (compact index)	High (due to large vector storage)
Interpretability	High (term-level match explanation)	Low (black-box embeddings)
Common Tools	BM25, TF-IDF, Elasticsearch	DPR
Suitability for Healthcare	Useful for structured queries and known terminology	Effective for unstructured clinical text and synonyms

doi:10.20944/preprints202508.1022.v1

3.2. RAG Variants

Several variants of RAG (see Figure 2) have advanced to balance trade-offs between training complexity, adaptability, and retrieval quality. These variants differ in how the retriever and generator interact, whether learning is coupled or decoupled, and how external knowledge is integrated into the generative process.

Figure 2. RAG Variants.

Naïve RAG employs a fixed, non-trainable retriever based on sparse methods (like TF-IDF or BM25) and a frozen LLM [10]. As illustrated in Figure 3, the pipeline begins with a user query, which is used to retrieve relevant documents from a pre-indexed corpus. The top-*k* retrieved passages are concatenated with the query to form an input prompt, which is passed to the frozen LLM. The model generates a response without receiving any feedback to improve the retriever, making the system static and suboptimal.

This approach is simple but suffers from limitations at multiple stages. It retrieves irrelevant or redundant content, struggles to synthesize fragmented information across documents, and generates responses that are unverified by the retrieved evidence due to a lack of explicit alignment between retrieval and generation.

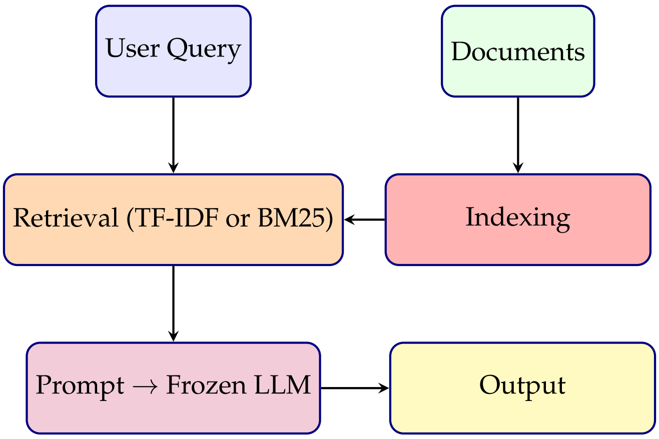

Figure 3. Naive RAG Architecture.

Advanced RAG (as illustrated in Figure 4) addresses the limitations of naïve RAG through a three-stage pipeline: pre-retrieval, retrieval, and post-retrieval enhancements [10]. The process begins with a user query and an indexed document corpus. During the pre-retrieval phase, query routing, rewriting, and expansion techniques improve retrieval quality by refining the input.

In the retrieval phase, semantically encoded queries are matched with indexed document chunks using fine-tuned embedding models and task-aligned similarity metrics (e.g., cosine similarity, dot product). Advanced strategies like hybrid retrieval (combining sparse and dense approaches) and prompt-adaptive conditioning further improve relevance.

Following retrieval, the post-retrieval module filters, re-ranks, and compresses retrieved chunks. Redundant or noisy results are removed using re-ranking or summarization techniques, ensuring only the most pertinent context reaches the frozen LLM for response generation. Despite no backpropagation to the retriever, this architecture significantly enhances factuality, semantic alignment, and clinical reliability over naïve approaches.

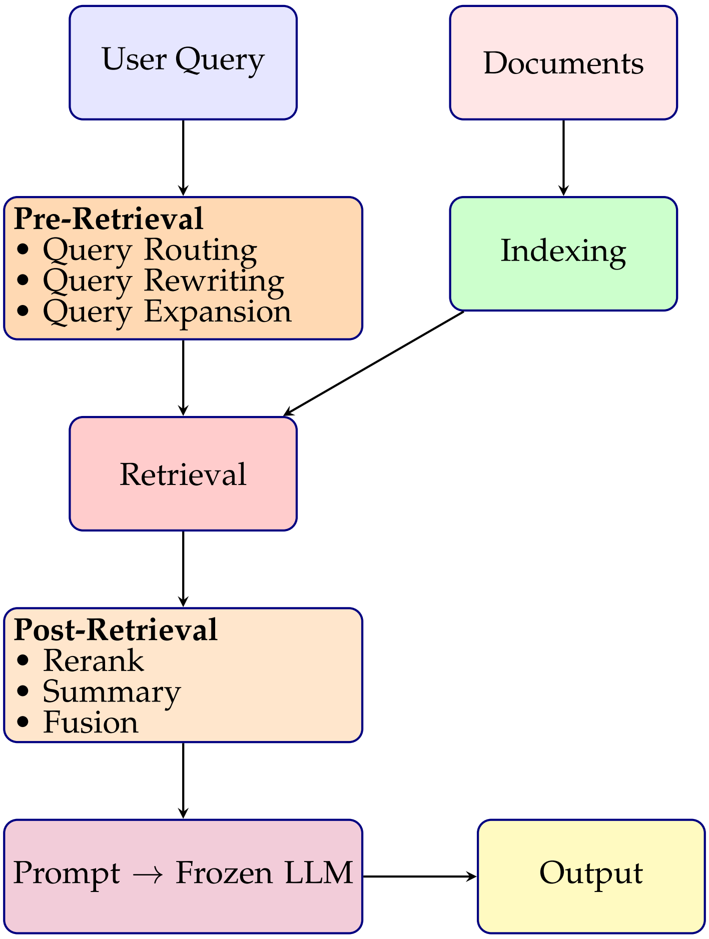

Figure 4. Advanced RAG Architecture.

Modular RAG extends the capabilities of advanced RAG by decomposing the architecture into independently configurable modules [23]. Unlike Naïve or Advanced RAG, Modular RAG separates the training of the retriever and generator. The retriever, a dense dual encoder, is trained using contrastive loss, and the generator is fine-tuned independently on top-k retrieved documents using maximum likelihood estimation. Although this improves modularity and flexibility, it limits retrieval adaptation during generation.

The architecture (see Figure 5) supports a variety of modules that can be selectively included or reconfigured to suit task-specific needs. These include: (1) a Search module that supplements retrieval with additional sources like web results or structured knowledge graphs; (2) a Memory module that maintains context from prior interactions, facilitating continuity in dialogue or multi-turn applications; and (3) Routing, which directs queries to the most suitable operations such as summarization or entity extraction. Modules like Rewrite refine queries before retrieval, Rerank ensure relevant results are prioritized, and Fusion combines outputs from multiple retrieval pathways to improve contextual alignment and reduce noise.

By enabling fine-grained control and extensibility, Modular RAG supports experimentation with novel retrieval strategies, integration of domain-specific knowledge, and adaptation of architectures for clinical, scientific, and enterprise use cases. It provides a unifying paradigm under which designs such as Naïve RAG, Advanced RAG, and iterative retrieval-generation loops can be interpreted as specialized module configurations.

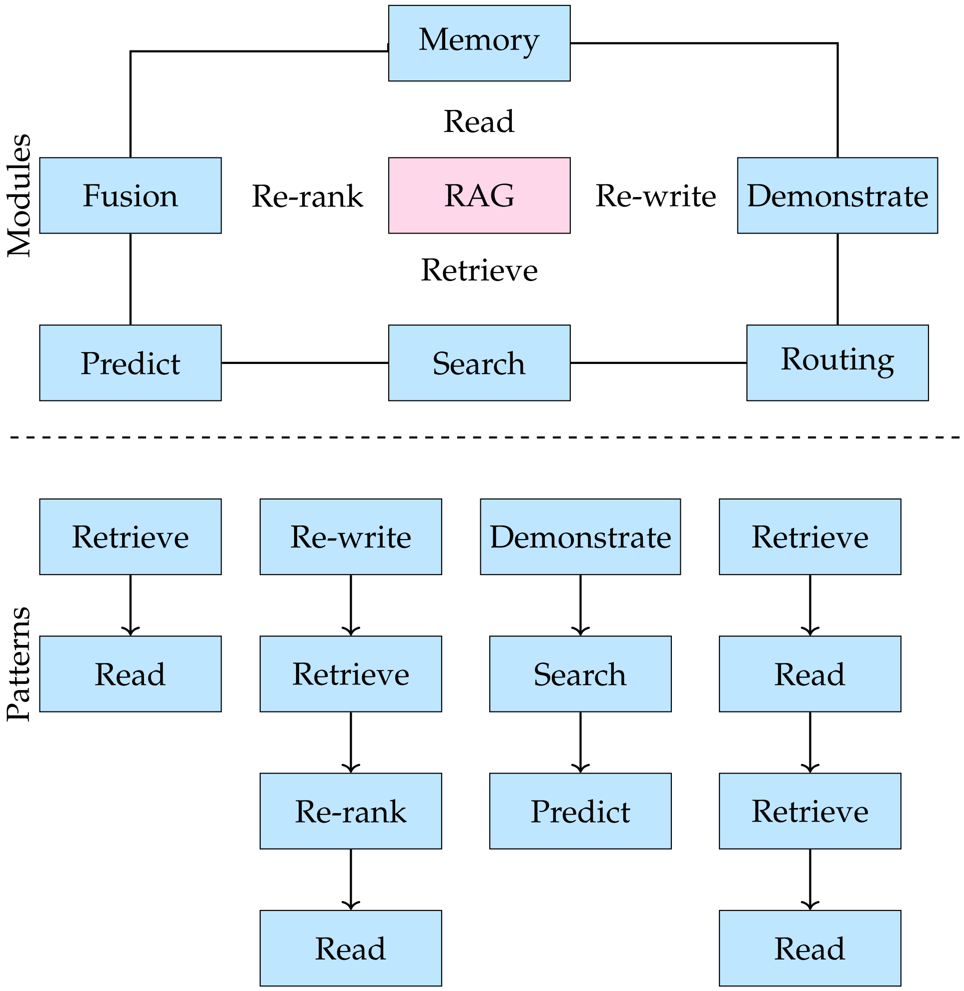

Figure 5. Modular RAG Architecture.

Table 4 provides a comparison of Naïve, Advanced, and Modular RAG Architectures.

Furthermore, RAG systems can also be classified based on their scope: (1) Open-domain RAG and (2) Domain-specific RAG.

Open-domain RAG retrieves information from broad, general-purpose corpora such as Wikipedia or web-scale datasets. Foundational architectures, including REALM [24], Atlas [25], and RETRO [26], fall into this category. These models differ in how retrieval is integrated: REALM emphasizes retriever pretraining with latent supervision; Atlas updates the retriever during fine-tuning; and RETRO performs chunk-level retrieval with local memory compression. While open-domain RAG enables wide coverage and strong generalization, it lacks the precision and factual reliability required in specialized domains such as healthcare.

Domain-specific RAG, by contrast, restricts retrieval to curated biomedical sources, enhancing factual grounding, terminological accuracy, and trustworthiness. Models such as MedRAG [27], BioRAG [28], and ClinicalGPT [29] exemplify this approach. These systems use biomedical embeddings (e.g., BioBERT, PubMedBERT [30]), specialized corpora (e.g., PubMed abstracts, clinical notes), and domain-specific knowledge bases (e.g., unified medical language system (UMLS), MIMIC). By customizing retrieval and generation to medical language, they enable reliable performance in tasks such as diagnostic decision support, clinical report summarization, patient-specific question answering, and medical dialogue generation.

Table 4. Comparison of Naïve, Advanced, and Modular RAG Architectures.

Keypoints	**Naïve RAG**	**Advanced RAG**	**Modular RAG**
Architecture	Simple two-stage pipeline: retrieval + generation	Three-stage pipeline: pre-retrieval, retrieval, post-retrieval	Fully decomposed pipeline with plug-and-play components
Query Processing	Uses raw user query	Query rewriting, expansion, or routing applied before retrieval	Modular query handling with flexible preprocessing units
Retriever Type	Dense retrievers (e.g., DPR)	Hybrid retrievers combining dense + sparse (e.g., BM25 + dense)	Modular and replaceable retrievers (dense, sparse, hybrid, trainable)
Post-Retrieval Handling	No reranking or filtering	Reranking, summarization, and filtering of retrieved chunks	Dedicated modules for reranking, deduplication, and compression
LLM Role	Frozen LLM processes retrieved documents directly	Frozen LLM with prompt-adaptive input conditioning	Swappable LLM head (frozen, fine-tuned, adapter-based)
Training Flexibility	No training of retriever or generator	Retriever may be fine-tuned; generator remains frozen	Independent or joint training of all modules (retriever, reranker, generator)
Transparency	Low interpretability; retrieval-to-generation is a black box	Some transparency with reranking scores or summarization	High transparency; traceable intermediate outputs for each module
Use Case Suitability	Basic Q&A and document retrieval tasks	High-stakes applications like medical QA, EHR summarization	Production-ready systems, customizable deployments, and MLOps integration
Latency	Low due to fewer stages	Moderate to high depending on pre/post processing complexity	Configurable latency depending on module choices
Customization	Minimal	Moderate pipeline-level customization	Full customization at component level

A comparative overview of RAG variants is provided in Table 5.

Table 5. Comparison of RAG Variants.

RAG Variant	**Training Strategy**	**Retriever Type**	**Advantages / Limitations**
Naïve	No training; static retrieval	Sparse (e.g., TF-IDF)	Simple; fast; no task alignment; poor factual grounding
Modular	Independent training of *R* and *G*	Dense dual encoders (e.g., DPR)	Modular, scalable; lacks feedback from generator
Advanced	Joint + feedback loops	Hybrid (dense + sparse), knowledge-enhanced	Factual, dynamic, and adaptable; complex to implement
Open-domain	Pretrained on general corpora	Generic (e.g., Wikipedia)	Broad scope; risks hallucination and low domain relevance
Domain-specific	Tuned on medical corpora	Biomedical (e.g., PubMed, MIMIC)	High clinical accuracy; limited generalization outside domain

4. Related Work

In recent years, research on the application of RAG-based LLMs in healthcare has grown substantially. To provide a structured understanding, this section organizes the related work into seven thematic areas: (1) general clinical support systems, (2) Chatbots for patient interaction, (3) specialty-specific implementations, (4) Signal and time-series analysis, (5) graph and ontology-enhanced RAG architectures, (6) privacy-preserving and secure RAG systems, and (7) radiology-focused RAG applications. Each subsection synthesizes representative studies, highlighting methodologies, key findings, and identified limitations.

4.1. General Clinical Applications of RAG

RAG-enhanced LLMs have been applied to general medical reasoning, discharge planning, and infectious disease support. These systems show effectiveness in outpatient care, triage, and nursing workflows by delivering accurate, context-aware responses that reduce clinician burden.

Kirubakaran et al. propose a RAG-based medical assistant for infectious disease support, integrating quantized LLMs, knowledge graphs, and speech synthesis [31]. The system utilizes graph databases for efficient context retrieval, XTTS (an open-source cross-lingual text-to-speech engine) for multilingual voice cloning, and combines naïve RAG, auto-merging, and ensemble retrievers to generate accurate responses. While effective in delivering COVID-19-related support, the system faces challenges related to multilingual scalability, real-time knowledge updates, and generating empathetic feedback.

Upadhyay et al. introduce a three-stage RAG-based model for health information retrieval that integrates PubMed-based passage extraction, LLM-generated text (GenText), and a dual scoring mechanism based on topical relevance (BM25) and factual accuracy (via stance detection and semantic similarity) [32]. The model outperforms existing baselines, particularly when using LLaMA, and improves explainability through citation-rich outputs that help users trace sources. However, limitations include dependence on general-purpose LLMs, approximate factuality assessment, and the risk of automation bias.

Yang et al. propose a lightweight RAG-based personalized discharge care system enhanced with a memory mechanism to support neurological patients post-hospitalization [33]. The system integrates long-term, short-term, and dynamic memory modules to personalize responses and employs prompt engineering alongside a custom vector database for efficient retrieval. Both GPT-4 and LLaMA 3 models were evaluated, with GPT-4 demonstrating superior performance; however, both benefited from memory-augmented RAG integration. The system is designed for scalability through vector and non-relational database architectures. However, limitations include evaluation bias, small-scale testing, lack of real-time clinical deployment, and the absence of robust alerting functionality.

Hammane et al. introduce a medical reasoning framework that enhances RAG with a self-evaluation mechanism to improve answer accuracy and clinical trustworthiness [34]. The system integrates external medical knowledge via retrieval, employs a policy model to generate candidate responses, and applies a reward model for automatic scoring and reranking. The key strengths include automated feedback and support for multi-stage reasoning. However, limitations remain in clinical validation, dependence on synthetic reward signals, and restricted evaluation beyond multiple-choice datasets.

Xu et al. evaluate a RAG-enhanced GPT system for breast cancer nursing care, showing improved response accuracy and nurse satisfaction compared to direct GPT outputs, while maintaining empathetic tone [35]. Using 15 randomized questions, senior nurses assessed responses across key qualitative metrics. However, the study is limited by the absence of patient validation, narrow data scope, lack of real-world deployment, and insufficient theoretical grounding for the observed performance gains.

Hsu et al. present a two-stage RAG-based system combining LLMs with clinical guidelines [36]. Unlike single-pass models, it first generates assessments from subjective and objective inputs, then produces personalized treatment plans using patient history and cross-patient data. Evaluation on real EHR data shows improved accuracy and relevance over baselines. However, the model's hospital-specific training limits generalizability, and it remains susceptible to LLM biases.

Aminan et al. introduce a glaucoma-specific RAG system using GPT-4.5 and a custom retrieval-response framework for diagnostic support [37]. It is built on curated clinical sources, and it outperforms GPT-4.5 and DeepSeek-R1 [1] in accuracy and relevance. However, its limitations include a lack of image input, retrieval noise, and a static knowledge base.

Thompson et al. propose a zero-shot phenotyping method using LLMs with retrieval-augmented generation (RAG) and MapReduce, which handles the large number of clinical text snippets retrieved via regular expressions, to identify diseases such as pulmonary hypertension from clinical notes [38]. Regular expression-based retrieval extracts relevant snippets, which are processed in parallel by an LLM, and the results are aggregated using max voting or prompting. The model outperforms rule-based baselines and generalizes across diverse note types. Its strengths include scalability and reduced manual effort, while limitations involve reliance on handcrafted regex rules, use of a single LLM, and lack of integration with structured EHR data.

Benfenati et al. propose a RAG-based question answering (QA) framework for nutrigenetics that combines general-purpose text embeddings, a vector database, and LLMs to enhance answer relevance without requiring model retraining [39]. The system uses curated abstracts for retrieval based on cosine similarity and augments prompts to improve accuracy and specificity. Evaluated on domain-specific questions, the RAG approach outperforms standard LLMs in terms of relevance, depth, and evidence support. However, limitations include dependence on abstracts, use of static context, and lack of support for multimodal data.

Ziletti et al. present a retrieval-augmented text-to-SQL framework for epidemiological QA using EHR and claims data [40]. It integrates medical coding and domain-specific RAG to improve SQL generation accuracy. Evaluated on a curated dataset across multiple LLMs, the method outperforms static prompting, with GPT-4 leading. While effective, limitations include small dataset size, narrow focus, and modest overall performance, indicating a need for broader datasets and more robust models.

Pyae et al. present a RAG-based system for personalized fitness and dietary guidance using wearable data and a curated knowledge base [41]. It integrates LLMs with voice and text interfaces. The system performs well in relevance and memory but lags in faithfulness due to hallucinations and limited data scope. Key gaps include weak voice recognition in noisy settings, lack of dynamic knowledge updates, and retrieval inaccuracy for complex queries.

Cheetirala et al. propose a RAG-based method for classifying surgical complications by selecting the most relevant 4,000 tokens from clinical notes, reducing computational cost without compromising accuracy [42]. The study compares standalone LLMs using RAG-based retrieval versus full-text ingestion, reporting comparable evaluation scores. Notably, token usage and inference costs were reduced by over 90% with RAG. While the approach is effective and scalable, limitations include static retrieval, a narrow task focus, and the absence of advanced retrieval techniques.

4.2. RAG Chatbots for Patient Interaction

RAG-based chatbots offer scalable solutions for patient self-management, pre-consult triage, and mental health screening, especially beneficial in telemedicine, rural care, and low-resource settings where clinician availability is limited.

Kulshreshtha et al. propose a RAG-based medical chatbot using standalone LLMs (LLaMa, LangChain, and BERT) to improve healthcare accessibility [43]. It uses context-aware embeddings, Faiss for retrieval, and LLM guard for Health Insurance Portability and Accountability Act (HIPAA)/ general data protection regulation (GDPR) compliance. The methodology include tokenization, retrieval, beam search, and post-processing for accurate, private responses. While effective in clinical support, there are gaps in multilingual support, bias reduction, and speech integration.

Shafi et al. present LLM-Therapist, a multimodal healthcare assistant combining RAG and real-time function calling for personalized behavioral support [44]. It uses patient data, vector-based semantic search, and GPT-3.5 with ChromaDB (open-source vector database) to generate context-aware responses. The key strengths include its hybrid architecture, use of function calling for real-time data, and domain-specific fine-tuning. However, it lacks large-scale clinical validation, benchmarking with other systems, and analysis of privacy, security, and deployment scalability.

Sree et al. introduce a RAG-based LLM chatbot for physical and mental healthcare, integrating external data sources to enhance diagnostic support and personalization [45]. It employs agents and modular tools for domain-specific retrieval and semantic understanding, using indexing and embedding-based search to generate real-time, relevant outputs. Evaluation shows improved latency, token efficiency, and user satisfaction. However, it lacks clinical validation, deployment benchmarks, error analysis, and considerations of privacy, data governance, and ethics, limiting practical applicability.

Sinha et al. propose a RAG-based chatbot for medical triage that integrates LLMs with vector embeddings and a retrieval system to generate efficient, context-aware responses [46]. The system achieves an average response time of 28 seconds and shows moderate performance on standard language generation metrics. It improves upon traditional chatbots through dynamic learning, voice input, and context retention. However, it lacks multi-turn robustness and has not been benchmarked against commercial or baseline systems, limiting its real-world applicability.

Nayinzira et al. introduce a mental health chatbot combining RAG and sentiment analysis [47]. Using a public counseling dataset, it evaluates Naive RAG, Multi-query RAG, and document embeddings with LLMs (GPT and Mistral). Results show sentiment analysis improves performance, especially with Multi-query RAG and Mistral, but increases latency and token use. Its key limitations are token limits and a lack of clinical data.

Shin et al. present a RAG-based chatbot for thyroid disease management that integrates ChatGPT-4o with a vector database built from guidelines and textbooks [48]. It produced more accurate, safe, and clinically applicable responses to patient-specific queries with fewer hallucinations as compared to standalone LLMs. However, its limitations include a small evaluator group and the absence of an ablation study.

4.3. Specialty-Focused RAG Models

RAG-based models show promise in specialty decision support, patient-specific education, and knowledge reinforcement in clinical subspecialties, making them suitable for many healthcare departmental deployment.

Miao et al. explores integrating RAG with LLMs to improve accuracy in nephrology [49]. It shows how a GPT-4-based system, aligned with KDIGO (Kidney Disease: Improving Global Outcomes) guidelines, outperforms standard GPT-4 by reducing hallucinations and enhancing clinical relevance. The methodology involves building a chronic kidney disease specific knowledge base, API integration, and fine-tuning. Applications include decision support, medical education, and multidisciplinary care. However, the work is conceptual, lacks clinical validation, and depends on data quality and retrieval accuracy.

Ge et al. developed a liver disease-specific LLM using RAG within a Protected Health Information (PHI)-compliant system [50]. They embedded 30 guidelines from the American Association for the Study of Liver Diseases (AASLD) using a text-embedding model and stored them in a vector database. User queries were embedded in real time, matched to the database, and processed by GPT-based LLMs. The model provided more specific answers than ChatGPT and correctly answered all benchmark questions, though some justifications were incomplete due to limited sources and contextual bias.

Long et al. present a domain-specific LLM designed for otolaryngology–head and neck surgery using Retrieval-Augmented Language Modeling [51]. By building a curated knowledge base and applying text-embedding for semantic search, the system augments queries with relevant domain content before passing them to ChatGPT. Evaluated on Canadian and US board-style questions, the model outperformed ChatGPT in validity, safety, and competency. However, gaps remain in coverage for areas like pediatrics and basic science.

4.4. RAG for Signal and Time-Series Tasks

RAG models using time-series data enhance diagnostic accuracy by integrating structured physiological signals into prompts. These systems support cardiology and pain management settings by aligning patient-generated data with clinical guidelines, enabling more personalized and guideline-consistent care.

Yu et al. propose a zero-shot retrieval-augmented LLM framework for ECG diagnosis, using domain-specific databases to guide prompt construction and inference [52]. Applied to arrhythmia and sleep apnea tasks, the method transforms ECG features into structured prompts enriched with expert knowledge. Evaluated using standalone LLMs (LLaMA2, GPT-3.5), the approach outperforms few-shot baselines and achieves performance comparable to supervised deep learning models. The key limitations include reliance on handcrafted features and susceptibility to signal noise.

Chen et al. introduce a clinical decision-support system for chronic low back pain that combines LLMs, RAG, and least-to-most prompting [53]. By integrating patient data and retrieving up-to-date clinical knowledge, it enhances accuracy and personalization. Compared to standalone LLMs, it scores highest in accuracy, relevance, clarity, benefit, and completeness. Strengths include dynamic knowledge grounding, incorporation of psychosocial factors, and prompt engineering for improved interpretability. However, limitations include the absence of long-term outcome validation and real-time knowledge updating.

4.5. Graph-Based and Ontology-Aware RAG Frameworks

Graph-enhanced RAG improves explainability and trust, making it particularly useful in patient education tools, informed consent generation, and clinical decision support where provenance and logic tracing are essential.

Rani et al. propose a graph-based RAG framework for diabetes care, addressing hallucination, explainability, and source traceability in LLMs [54]. It integrates keyword, vector, and graph retrieval with a validated knowledge graph using topic-based chunking and semantic matching. The model requires improvement in long-answer summarization and broader applicability across medical domains.

Wu et al. introduce a graph-based RAG method for medical question answering [55]. It improves reliability by linking user data to trusted medical sources using Triple Graph Construction and retrieves answers through a two-step U-Retrieval process. Tested on benchmarks, it outperforms standard RAG, GraphRAG [56], and fine-tuned medical LLMs. However, challenges remain in adapting to live data, complex patient queries, and maintaining explainability.

Sophaken et al. present a Graph-RAG pipeline for dentistry that integrates Named Entity Recognition (NER), Resource Description Framework (RDF) triples, and RAG to support clinical decision-making [57]. Multimodal data is processed using LLMs to extract entities and relations, structured as RDF triples and embedded into a dental ontology. This enables both vector and graph-based retrieval. Experimental results show improved precision and recall through reranking, with triple store RAG enhancing diagnostic accuracy. However, the system depends on complete ontologies, is computationally intensive, and is currently limited in clinical scope.

Shi et al. introduce a retrieval-augmented LLM for shared decision-making in adolescent idiopathic scoliosis (AIS), combining ChatGPT with a structured AIS-specific knowledge base and dense retrieval using MPNet (Masked and Permuted Pre-training for Language Understanding) to enhance response accuracy, relevance, and transparency [58]. Evaluations through usability tests, knowledge assessments, and expert reviews show improved patient understanding and superior performance over ChatGPT. The model reduces hallucinations through source-grounded reasoning enabled by ReAct (Reasoning and Acting) prompting. The limitations include a static, keyword-filtered knowledge base, a lack of multimodal inputs, and limited clinical validation.

doi:10.20944/preprints202508.1022.v1

4.6. RAG with Blockchain and Secure Architectures

RAG models have also been extended with blockchain and smart contract mechanisms to ensure secure retrieval and auditability in sensitive domains such as pediatrics and Internet of Medical Things (IoMT). These efforts aim to enhance trustworthiness and decentralize data governance.

Su et al. propose a Hybrid RAG-based Multimodal LLM framework for secure IoMT data management, integrating RAG with a diffusion-based smart contract for explainable and trustworthy retrieval [59]. It combines semantic search, multimodal inputs, and blockchain to ensure secure, contextually relevant access to medical data. This approach lacks scalability evaluation and comparative analysis with other contract methods.

Jabarulla et al. present a blockchain-enabled RAG system for pediatric clinical decision support, combining semantic search over guideline PDFs with LLMs to generate responses, which clinicians evaluate and store via smart contracts on a permissioned blockchain [60]. It ensures secure, auditable feedback for model refinement but lacks large-scale deployment, diverse validation, real-world hospital integration, and automated fine-tuning, limiting clinical scalability.

4.7. Radiology-Specific Retrieval-Augmented QA

RAG-based tools can be integrated into radiologists' workstations to assist with second opinions, structured reporting, and continuous medical education, particularly in complex imaging cases.

Arasteh et al. proposed a RAG system for radiology QA that retrieves context from Radiopaedia [61]. Unlike static RAG setups, it dynamically fetches relevant articles, embeds them, and uses them to guide LLMs. Tested on benchmark datasets, RadioRAG gave improved accuracy for open-source models, while reducing hallucinations and improving factuality. However, gains varied by model, and strict reliance on retrieved context occasionally led to errors. Limitations include dependence on a single source, a small external dataset, and slower response times.

5. Applications

This section outlines the key applications of RAG in healthcare, emphasizing clinical utility, factual consistency, and reliability.

5.1. Diagnostic Assistance

RAG enhances diagnostic decision-making by retrieving similar patient cases or relevant clinical guidelines from structured databases and unstructured literature. In radiology and pathology, image captions or structured reports are used to query repositories, enabling the generation of differential diagnoses or evidence-backed recommendations. This retrieval-grounded approach improves factual accuracy and reduces dependence on parametric memory.

5.2. Summarization of EHRs and Discharge Notes

RAG supports clinical summarization by integrating longitudinal patient data. The retriever identifies pertinent clinical history such as medications, lab trends, and prior admissions from EHRs, allowing the generator to produce concise discharge summaries or progress notes. Both extractive and abstractive outputs benefit from evidence-based contextualization, minimizing omissions and clinical inaccuracies. Dense retrievers using domain-tuned embeddings (e.g., BioBERT) yield superior semantic alignment over sparse methods.

5.3. Medical Question Answering

RAG-based QA systems address complex clinical queries by retrieving content from curated biomedical corpora (e.g., PubMedQA). By anchoring answers in peer-reviewed literature or structured medical knowledge, these systems reduce hallucinations often seen in standalone LLMs. Indexing strategies like BM25 ensure relevance in long-form or open-ended queries, improving both reliability and transparency.

5.4. Patient Education and Conversational Agents

RAG-powered conversational agents enhance patient engagement by retrieving validated educational material and tailoring responses to user profiles. This ensures that medical concepts are accurately translated into layperson language without compromising factual integrity. Retrieval from authoritative sources reduces misinformation risks, making RAG suitable for triage assistants and post-discharge counseling tools.

5.5. Clinical Trial Matching

RAG enables precise clinical trial matching by aligning patient profiles with eligibility criteria extracted from trial registries. The retriever identifies suitable protocols based on demographics, comorbidities, and biomarkers, while the generator summarizes inclusion/exclusion criteria in natural language. This improves trial enrollment and expands therapeutic access, particularly in rare or refractory cases.

5.6. Biomedical Literature Synthesis

RAG facilitates dynamic literature review by retrieving and summarizing recent, domain-relevant publications in response to specific clinical or research queries. Instead of manually scanning large corpora, clinicians and researchers receive context-aware summaries grounded in high-quality sources (e.g., PubMed, PMC). This streamlines evidence-based practice and reduces cognitive burden during decision-making or guideline development.

6. Evaluation Framework: Metrics and Benchmarks

The evaluation of RAG models in healthcare requires a multidimensional approach, focusing on generation quality, retrieval effectiveness, factual consistency, and clinical validity.

6.1. Domain-Specific Metrics for Clinical Validation

- **FactScore** assesses factual alignment between generated outputs and reference data, particularly for medical summaries and treatment plans [62].
- **RadGraph-F1** measures overlap between generated and reference entity-relation graphs in radiology reports, ensuring structural and factual correctness [63].
- **MED-F1** quantifies alignment of extracted entities with standard clinical terminology and is frequently applied to tasks like medical named entity recognition, particularly in datasets such as CheXpert [64].

These metrics are used in medical contexts, capturing domain-specific vocabulary, reasoning patterns, and structured knowledge representations that generic metrics fail to address.

6.2. Generation Quality Metrics

Several standard natural language generation (NLG) metrics are employed to assess the textual quality of model outputs, including:

- **BLEU (Bilingual Evaluation Understudy)** measures the degree of n-gram overlap between the generated and reference text. It is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

  where p_n denotes modified n-gram precision, w_n are weights (commonly uniform), and BP is the brevity penalty to discourage overly short outputs.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**, particularly ROUGE-L, evaluates the longest common subsequence between generated and reference texts, and is frequently used in summarization tasks involving clinical documents.

- **F1 Score** is the harmonic mean of precision and recall, particularly relevant for span-based extraction and classification tasks.
- **BERTScore** compares contextual token embeddings between candidate and reference texts using models such as BioBERT, offering semantic alignment beyond surface-level matching.

While these metrics evaluate fluency and coherence, they fall short in assessing factual correctness in high-stakes clinical applications.

6.3. Retrieval Relevance Metrics

The retrieval component in RAG systems is assessed by evaluating the relevance and ranking of retrieved documents. Common metrics include:

- **Recall@k** measures the fraction of relevant documents retrieved within the top-k results:

$$\text{Recall@}k = \frac{|\text{Relevant} \cap \text{Retrieved@}k|}{|\text{Relevant}|}$$

- **Mean Reciprocal Rank (MRR)** evaluates the average inverse rank of the first relevant document:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

- **Normalized Discounted Cumulative Gain (nDCG@k)** considers both the relevance and position of retrieved documents:

$$\text{nDCG@}k = \frac{DCG@k}{IDCG@k}, \quad \text{where } DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$

These metrics ensure that relevant clinical evidence is retrieved and appropriately ranked, directly impacting the reliability of the generated outputs.

6.4. Factual Consistency Metrics

Ensuring factual alignment between generated outputs and source data is critical in healthcare applications:

- **FEVER Score** quantifies the correctness of generated claims against retrieved evidence. It has been adapted from open-domain fact-checking to biomedical domains.
- **Faithfulness Metrics** evaluate the consistency of generated responses with retrieved or reference content, using either entailment-based models or domain-specific factuality checkers.

Standard metrics such as BLEU, ROUGE, and BERTScore primarily focus on lexical or semantic similarity but are insufficient for validating factual correctness in high-stakes medical contexts.

Additional practical metrics include:

- **Response Time (Latency)**: Critical for real-time applications such as triage or bedside decision support, where generation delay can impact clinical workflow.
- **Source Traceability**: Refers to the model's ability to link generated content back to specific retrieved sources, thereby enhancing transparency, auditability, and user trust.

6.5. Benchmark Datasets

The following datasets are widely utilized for evaluating RAG systems in healthcare contexts:

- **MedQA (USMLE)**: A multiple-choice question dataset derived from medical licensing exams, focused on clinical knowledge assessment [65].
- **PubMedQA**: Consists of biomedical abstracts paired with yes/no/maybe questions, requiring grounded reasoning and evidence-based answers [66].

- **MIMIC-IV**: A comprehensive, de-identified EHR dataset supporting tasks such as summarization, question answering, and document retrieval [17].
- **MedDialog**: A multilingual dataset of doctor-patient conversations, suitable for training and evaluating medical dialogue systems [67].

These benchmarks collectively support a comprehensive and reproducible evaluation of RAG models deployed in medical domains.

7. Challenges and Limitations

Despite its promise, RAG faces several technical and practical challenges in healthcare applications. The key issues include retrieval accuracy, generation latency, limited explainability, and data privacy concerns.

Table 6 summarizes the key challenges in deploying RAG systems in healthcare, outlining their underlying causes, consequences, and potential mitigation strategies.

Table 6. Challenges in RAG Deployment in Healthcare.

Challenge	Cause	Consequence	Mitigation
Domain Shift & Retrieval Noise	Heterogeneous EHR styles, outdated or mixed-quality sources	Retrieval mismatch, irrelevant or unsafe generations	Domain-adaptive retrievers, curated clinical corpus, context filters
Latency	Sequential retrieval and generation over large corpora	Delayed responses in real-time clinical scenarios	Lightweight retrievers, caching, on-device or edge retrieval
Lack of Explainability	no attribution linking sources to generated content	Low clinician trust, limited transparency	Source highlighting, rationale extraction, evidence traceability
Privacy & Compliance Risks	Inadequate de-identification, unrestricted protected health information access	Legal violations, re-identification risk	Secure indexing, redaction, audit trails, access control
Weak Clinical Retrieval	General retrievers overlook domain-specific semantics	Missed context, hallucinated content	Biomedical retrievers (e.g., BioBERT), ontology-guided search (UMLS)
Noisy & Unstructured Clinical Text	Abbreviations, typos, incomplete or inconsistent notes	Imprecise embeddings, factual drift	Preprocessing pipelines, clinical QA models, structured input templates
Evaluation Limitations	Generic NLP metrics, lack of clinical gold standards	Poor assessment of safety, factuality, and utility	Domain-specific metrics (FactScore, MED-F1), expert-in-the-loop evaluations
Multimodal Limitations	Text-only retrieval ignores imaging, labs, genomics	Incomplete or narrow decision support	Multimodal encoders, joint indexing, cross-modal retrieval
Infrastructure Constraints	High storage/compute requirements, poor connectivity	Limited feasibility in low-resource settings	Model compression, retriever distillation, offline retrieval setups
Knowledge Drift	Static models and outdated retrieval indices	Obsolete or harmful recommendations	Continual learning, live corpus updates, dynamic retrievers
Lack of Human Oversight	Fully automated pipelines without expert feedback	Errors propagate unchecked, especially in diagnosis	Feedback interfaces, clinician-in-the-loop retrieval and validation
Bias and Fairness	Skewed training corpora, underrepresented populations	Health disparities, biased or unsafe outputs	Diverse data curation, fairness evaluation, inclusive retriever tuning

7.1. Retrieval Challenges in Clinical Contexts

Healthcare data is highly heterogeneous, varying across institutions, patient populations, documentation styles, and EHR systems. This heterogeneity introduces domain shift, where the data distribution during inference diverges from the training or retrieval corpus. Such shifts impair retriever performance and hinder the generator's capacity to produce clinically valid outputs. Mismatches stem from divergent medical terminologies, regional practices, or coding standards.

Additionally, clinical corpora blend curated medical literature with layperson articles, outdated sources, or non-standard documents. This introduces retrieval noise, increasing the risk of injecting factually incorrect or irrelevant context into model outputs. Further compounding the problem, generic retrievers struggle to capture biomedical semantics, failing to align domain-specific synonymy (e.g., "myocardial infarction" vs. "heart attack") and hierarchical medical taxonomies. Without fine-tuning on clinical corpora or integration with ontologies such as UMLS, retrieval results are frequently semantically misaligned.

Moreover, clinical text with shorthand, incomplete sentences, and non-standard abbreviations further degrades retriever performance. Such unstructured documentation confuses embedding models and leads to context mismatch, factual drift, and loss of temporal or causal nuance in generation. Effective retrieval in healthcare requires domain-adaptive retrievers, curated corpora, and preprocessing strategies to clean, normalize, and structure clinical text.

7.2. Latency and Real-Time Applicability

RAG pipelines operate in sequential stages such as retrieval, encoding, and generation, each introducing latency. In high-acuity environments like emergency rooms or intensive care units, even minimal delays compromise clinical usability. Latency is exacerbated by large corpora, complex embedding models, and non-optimized infrastructure. Network dependencies, inefficient input/output, and absence of caching strategies further hinder responsiveness. Real-time applicability demands latency-aware optimizations such as on-device retrieval, approximate search, or lightweight retrievers.

7.3. Explainability and Source Attribution

While RAG improves factuality through retrieval, the generation process remains opaque. It lacks explicit attribution or source-linked rationales, making it difficult for clinicians to trace the origin of recommendations or validate generated insights. This lack of transparency undermines trust, especially in clinical decision-making. Enhancing explainability requires source highlighting, rationale extraction, and citation tracing mechanisms to ensure traceability and support informed clinical oversight.

7.4. Privacy, Compliance, and Governance

RAG models operating over clinical data must comply with stringent privacy regulations (e.g., HIPAA, GDPR) and institutional review protocols. Even minimal leakage through metadata, misidentified information, or inference, can pose re-identification risks. Dense retrievers lacking proper access control may inadvertently surface protected information. Challenges compound in cross-institutional deployments involving federated corpora, consent management, and audit logging. Ensuring secure indexing, retrieval, filtering, and context redaction is essential for ethically responsible deployment.

7.5. Evaluation Bottlenecks in Clinical Contexts

Traditional NLP metrics (e.g., BLEU, ROUGE, BERTScore) emphasize surface similarity and fail to capture clinical correctness, safety, or reasoning quality. They overlook hallucinations, omissions, and domain-specific inaccuracies. Domain-aligned metrics like FactScore and RadGraph-F1 offer partial improvements but still lack the granularity needed to assess clinical reasoning or contextual fidelity. Human expert evaluations remain the gold standard but are expensive, non-scalable, and difficult to reproduce. The absence of robust, scalable, and clinically meaningful benchmarks is a critical barrier to reliable RAG evaluation.

7.6. Multimodal Limitations

Healthcare decision-making relies on diverse modalities, including radiology images, lab values, pathology reports, and genomics. Current RAG systems, primarily text-centric, cannot natively process multimodal inputs. Cross-modal retrieval and embedding frameworks remain underdeveloped, and the lack of joint indexing and fusion strategies restricts comprehensive clinical insight generation. Advancing RAG in healthcare necessitates foundational research in multimodal alignment, cross-attention mechanisms, and unified retrieval architectures.

7.7. Infrastructure and Scalability Constraints

RAG deployment in resource-constrained settings such as rural hospitals or low-income regions is challenged by hardware limitations (e.g., limited GPUs), large storage demands for vector indices, and unstable internet connectivity. The compute-heavy nature of dense retrievers and large LLMs restricts feasibility. Techniques like retrieval compression, knowledge distillation, and edge deployment remain underexplored but are essential to scale RAG equitably across healthcare systems.

7.8. Continual Learning and Knowledge Drift

Static RAG systems quickly become outdated as medical knowledge evolves. Without automated mechanisms for incorporating new research findings, clinical guidelines, or drug approvals, these systems risk perpetuating obsolete or harmful recommendations. Manual updates are slow and error-prone. Incorporating continual learning via retriever re-indexing, incremental fine-tuning, or streaming updates—is crucial to maintain relevance and clinical reliability over time.

7.9. Lack of Human-in-the-Loop

Most RAG models operate autonomously without clinician oversight. This limits accountability, especially in high-stakes scenarios such as differential diagnosis or treatment recommendation. Without human validation, errors can propagate unchecked, compromising patient safety. Incorporating mechanisms for clinician feedback, correction, and control during retrieval and generation phases is essential to ensure transparency, foster trust, and align outputs with real-world clinical judgment.

7.10. Bias and Fairness Concerns

RAG systems trained on biased corpora or skewed demographic distributions risk perpetuating disparities in care. Under-representation of minority populations in training data can lead to diagnostic inaccuracies, exclusion from clinical recommendations, or harmful stereotypes. Biases in retrieval can amplify these effects by selectively surfacing unbalanced or inappropriate content. Ensuring fairness requires curated, diverse corpora and bias-mitigation strategies during both the retrieval and generation phases.

8. Discussion and Future Directions

Advancing RAG for healthcare requires focused research on technical innovations, integration of structured medical knowledge, and improvements in safety, scalability, and usability. This section outlines key future directions to enhance the clinical applicability and reliability of RAG systems.

8.1. RAG with Knowledge Graphs

Incorporating structured biomedical knowledge into RAG systems, as exemplified by GraphRAG, enhances retrieval precision and factual grounding through semantic alignment between queries and domain-specific entities. Knowledge graphs enable relation inference and filtering of irrelevant content, improving clinical relevance. However, most existing systems rely on static graphs and lack support for temporal, hierarchical, or causal reasoning. Future directions include dynamic graph construction, integration of diverse medical ontologies, and use of graph neural networks to support real-time, scalable, and context-aware retrieval. Evaluating their impact on safety, hallucination mitigation, and interpretability remains essential for clinical deployment.

8.2. Continual Learning and Dynamic Retrieval

Static RAG systems fail to reflect evolving medical guidelines and discoveries. Updating retrieval indices regularly and fine-tuning models on new clinical data can maintain relevance and reduce factual drift over time. Incorporating streaming data pipelines, active learning loops, and low-overhead reindexing mechanisms enables timely adaptation to emerging clinical knowledge. Dynamic retrievers that adjust based on usage patterns or expert feedback further enhance reliability in changing clinical contexts.

8.3. Multimodal Integration

Clinical decisions rely on more than just text. Extending RAG systems to jointly process radiology images, pathology slides, lab reports, and structured EHRs can improve accuracy and expand use cases beyond text-based diagnosis. Multimodal RAG architectures can align features from different data types using cross-modal encoders or fusion techniques, enabling retrieval conditioned on both visual and textual cues. This supports tasks like image-report generation, pathology classification, and integrated decision support, where isolated text-based reasoning is insufficient.

8.4. Federated and Privacy-Preserving RAG

Future RAG frameworks must support federated training, differential privacy, and encrypted retrieval to protect patient data. This allows knowledge sharing across institutions without compromising data confidentiality. Federated RAG enables retriever and generator updates across decentralized datasets without central aggregation. Techniques like homomorphic encryption and secure multiparty computation ensure that sensitive information is never exposed during training or retrieval. Differential privacy mechanisms can further limit data leakage by adding noise, preserving utility while ensuring compliance with HIPAA, GDPR, and other regulations.

8.5. Task-Specific Evaluation Frameworks

General NLP metrics do not capture clinical correctness or safety. Domain-specific benchmarks, simulated patient scenarios, and expert-reviewed datasets are needed to evaluate medical RAG systems reliably. Evaluation must consider factual consistency, clinical validity, and potential harm. Metrics should assess alignment with guidelines, completeness of retrieved context, and appropriateness of generated responses. Developing standardized datasets with expert annotations and failure cases can help benchmark performance across specialties. Automated clinical validators and utility scoring models can further streamline scalable, safety-aware evaluation.

8.6. Human-in-the-Loop RAG Systems

Including clinicians in the retrieval and response validation loop enhances output quality and supports trust. Interactive interfaces can allow refinement of queries, re-ranking of evidence, and correction of generation errors. Such systems facilitate collaborative decision-making by enabling domain experts to validate retrieved context, flag hallucinations, and provide corrective feedback. Real-time user input can be logged for adaptive retriever tuning or future audits. Incorporating human oversight reduces automation bias, improves accountability, and ensures that generated responses align with clinical intent.

8.7. RAG for Low-Resource Settings

In many regions, clinical infrastructure is limited. Developing lightweight RAG models that work offline or with minimal compute, using localized or compressed medical corpora, can expand access to AI-based support. Optimization techniques such as model quantization, knowledge distillation, and efficient retrieval algorithms can reduce hardware demands. Offline-compatible RAG systems, preloaded with region-specific content, ensure continuity of care in areas with unreliable internet. Additionally, tailoring language models to local clinical terminology and cultural contexts improves usability and relevance.

8.8. Explainable RAG Pipelines

Traceability from generated output back to retrieved sources is critical in healthcare. Visualizing evidence contributions and highlighting decision paths can improve clinician confidence and support audits. Methods such as token-level attribution, retrieval relevance scoring, and citation-style references help clarify the link between retrieved chunks and generated responses. Embedding this transparency within clinical decision tools promotes accountability and facilitates regulatory compliance.

8.9. Clinical Workflow Integration

Standalone RAG tools are unlikely to be adopted. Embedding RAG into existing EHR interfaces and decision support dashboards with minimal disruption to clinical routines is essential for real-world use. Seamless integration ensures clinical relevance, reduces user friction, and aligns with existing workflows. APIs and middleware can enable real-time retrieval and generation within clinician-facing software, while maintaining compatibility with institutional IT infrastructure and compliance standards.

8.10. Bias Mitigation and Fairness Audits

Biases in training data or retrieval corpora can lead to uneven care recommendations. Ongoing audits and inclusion of diverse patient data during development can help ensure equitable performance across demographics. Techniques such as bias detection, reweighting, and fairness-aware retrieval can mitigate disparities, while evaluation across sub-populations ensures clinical safety and inclusivity.

9. Conclusion

RAG represents a significant advancement in aligning large language models (LLMs) with clinical knowledge needs by enabling dynamic retrieval, improving factual accuracy, and enhancing decision support. This review synthesized 30 peer-reviewed studies, categorizing RAG architectural variants, examining their deployment across diagnostic support, EHR summarization, and medical question answering, comparing standard and clinical-specific evaluation metrics, and identifying persistent deployment challenges in healthcare contexts.

Several limitations of this review must be acknowledged. The evolving terminology surrounding RAG may have led to the inadvertent exclusion of relevant studies. The analysis relied solely on reported findings without access to raw experimental data, and the review was bounded by a restricted publication window. These factors may limit completeness and generalizability.

Beyond the scope of this review, critical challenges remain for practical RAG deployment, including institutional resistance to AI adoption, regulatory alignment, interoperability with existing EHR systems, sustaining model performance amid domain shift, and establishing clinician trust. Addressing these barriers will require more resilient and adaptable RAG architectures.

Future research should prioritize continual learning, multimodal integration, standardized clinical evaluation protocols, and human-in-the-loop designs. Progress toward scalable, interpretable, and trustworthy RAG systems will be essential for enabling safe, equitable, and sustainable adoption in dynamic healthcare environments, benefiting researchers seeking robust methods, clinicians aiming for reliable decision support, and policymakers shaping responsible AI integration.

References

1. Neha, F.; Bhati, D. A Survey of DeepSeek Models. *Authorea Preprints*.
2. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nature Medicine* **2025**, *31*, 943–950.
3. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* **2022**, *23*, bbac409.

4. Chow, J.C.; Li, K. Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks. *Information* **2025**.
5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
6. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240.
7. Huang, K.; Altosaar, J.; Ranganath, R. ClinicalBert: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv preprint arXiv:1904.05342* **2019**.
8. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **2020**, *33*, 9459–9474.
9. Ng, K.K.Y.; Matsuba, I.; Zhang, P.C. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. *Nejm Ai* **2025**, *2*, AIra2400380.
10. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* **2023**, *2*.
11. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-Augmented Generation for AI-Generated Content: A Survey, 2024, [arXiv:cs.CV/2402.19473].
12. Gaur, M. Knowledge-Infused Learning. PhD thesis, University of South Carolina, 2022.
13. Spasic, I.; Nenadic, G. Clinical text data in machine learning: systematic review. *JMIR medical informatics* **2020**, *8*, e17984.
14. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena* **2020**, *404*, 132306.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
16. Kotei, E.; Thirunavukarasu, R. A systematic review of transformer-based pre-trained language models through self-supervised learning. *Information* **2023**, *14*, 187.
17. Johnson, A.E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* **2023**, *10*, 1.
18. Lu, Q.; Dou, D.; Nguyen, T. ClinicalT5: A generative language model for clinical text. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 5436–5443.
19. Bafna, P.; Pramod, D.; Vaidya, A. Document clustering: TF-IDF approach. In Proceedings of the 2016 International conference on electrical, electronics, and optimization techniques (ICEEOT). IEEE, 2016, pp. 61–66.
20. Amati, G. BM25. In *Encyclopedia of database systems*; Springer, 2009; pp. 257–260.
21. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.S.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the EMNLP (1), 2020, pp. 6769–6781.
22. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **2020**, *21*, 1–67.
23. Gao, Y.; Xiong, Y.; Wang, M.; Wang, H. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks. *arXiv preprint arXiv:2407.21059* **2024**.
24. Zhu, Y.; Ren, C.; Xie, S.; Liu, S.; Ji, H.; Wang, Z.; Sun, T.; He, L.; Li, Z.; Zhu, X.; et al. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016* **2024**.
25. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* **2023**, *24*, 1–43.
26. Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G.B.; Lespiau, J.B.; Damoc, B.; Clark, A.; et al. Improving language models by retrieving from trillions of tokens. In Proceedings of the International conference on machine learning. PMLR, 2022, pp. 2206–2240.
27. Zhao, X.; Liu, S.; Yang, S.Y.; Miao, C. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In Proceedings of the Proceedings of the ACM on Web Conference 2025, 2025, pp. 4442–4457.

28. Wang, C.; Long, Q.; Xiao, M.; Cai, X.; Wu, C.; Meng, Z.; Wang, X.; Zhou, Y. Biorag: A rag-llm framework for biological question reasoning. *arXiv preprint arXiv:2408.01107* **2024**.
29. Wang, G.; Yang, G.; Du, Z.; Fan, L.; Li, X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968* **2023**.
30. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **2021**, *3*, 1–23.
31. S, S.K.; G, J.W.K.; E, G.M.K.; J, M.R.; Singh A, R.G.; E, Y. A RAG-based Medical Assistant Especially for Infectious Diseases. In Proceedings of the 2024 International Conference on Inventive Computation Technologies (ICICT), 2024, pp. 1128–1133. https://doi.org/10.1109/ICICT60155.2024.10544639.
32. Upadhyay, R.; Viviani, M. Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy. *Discover Computing* **2025**, *28*, 27.
33. Yang, Y.; Xu, C.; Guo, J.; Feng, T.; Ruan, C. Improving the RAG-based personalized discharge care system by introducing the memory mechanism. In Proceedings of the 2025 IEEE 17th International Conference on Computer Research and Development (ICCRD). IEEE, 2025, pp. 316–322.
34. Hammane, Z.; Ben-Bouazza, F.E.; Fennan, A. SelfRewardRAG: enhancing medical reasoning with retrieval-augmented generation and self-evaluation in large language models. In Proceedings of the 2024 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE, 2024, pp. 1–8.
35. Xu, R.; Hong, Y.; Zhang, F.; Xu, H. Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses. *Scientific Reports* **2024**, *14*, 30794.
36. Hsu, H.L.; Dao, C.T.; Wang, L.; Shuai, Z.; Phan, T.N.M.; Ding, J.E.; Liao, C.C.; Hu, P.; Han, X.; Hsu, C.H.; et al. MEDPLAN: A Two-Stage RAG-Based System for Personalized Medical Plan Generation. *arXiv preprint arXiv:2503.17900* **2025**.
37. Aminan, M.I.; DARNELL, S.S.; Delsoz, M.; Nabavi, S.A.; Wright, C.; Kanner, E.; Jerkins, B.; Yousefi, S. GlaucoRAG: A Retrieval-Augmented Large Language Model for Expert-Level Glaucoma Assessment. *medRxiv* **2025**, pp. 2025–07.
38. Thompson, W.E.; Vidmar, D.M.; Freitas, J.K.D.; Pfeifer, J.M.; Fornwalt, B.K.; Chen, R.; Altay, G.; Manghnani, K.; Nelsen, A.C.; Morland, K.; et al. Large Language Models with Retrieval-Augmented Generation for Zero-Shot Disease Phenotyping, 2023, [arXiv:cs.AI/2312.06457].
39. Benfenati, D.; De Filippis, G.M.; Rinaldi, A.M.; Russo, C.; Tommasino, C. A retrieval-augmented generation application for question-answering in nutrigenetics domain. *Procedia Computer Science* **2024**, *246*, 586–595.
40. Ziletti, A.; DAmbrosi, L. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. In Proceedings of the Proceedings of the 6th Clinical Natural Language Processing Workshop; Naumann, T.; Ben Abacha, A.; Bethard, S.; Roberts, K.; Bitterman, D., Eds., Mexico City, Mexico, 2024; pp. 47–53. https://doi.org/10.18653/v1/2024.clinicalnlp-1.4.
41. Pyae, M.S.; Phyo, S.S.; Kyaw, S.T.M.M.; Lin, T.S.; Chondamrongkul, N. Developing a RAG Agent for Personalized Fitness and Dietary Guidance. In Proceedings of the 2025 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON). IEEE, 2025, pp. 600–605.
42. Cheetirala, S.N.; Raut, G.; Patel, D.; Sanatana, F.; Freeman, R.; Levin, M.A.; Nadkarni, G.N.; Dawkins, O.; Miller, R.; Steinhagen, R.M.; et al. Less Context, Same Performance: A RAG Framework for Resource-Efficient LLM-Based Clinical NLP. *arXiv preprint arXiv:2505.20320* **2025**.
43. Kulshreshtha, A.; Choudhary, A.; Taneja, T.; Verma, S. Enhancing Healthcare Accessibility: A RAG-Based Medical Chatbot Using Transformer Models. In Proceedings of the 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD). IEEE, 2025, pp. 1–4.
44. Shafi, F.R.; Hossain, M.A. Llm-therapist: A rag-based multimodal behavioral therapist as healthcare assistant. In Proceedings of the GLOBECOM 2024-2024 IEEE Global Communications Conference. IEEE, 2024, pp. 2129–2134.
45. Sree, Y.B.; Sathvik, A.; Akshit, D.S.H.; Kumar, O.; Rao, B.S.P. Retrieval-augmented generation based large language model chatbot for improving diagnosis for physical and mental health. In Proceedings of the 2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE). IEEE, 2024, pp. 1–8.
46. Sinha, K.; Singh, V.; Vishnoi, A.; Madan, P.; Shukla, Y. Healthcare Diagnostic RAG-Based Chatbot Triage Enabled by BioMistral-7B. In Proceedings of the 2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN). IEEE, 2024, pp. 333–338.

47. Nayinzira, J.P.; Adda, M. SentimentCareBot: Retrieval-augmented generation chatbot for mental health support with sentiment analysis. *Procedia Computer Science* **2024**, *251*, 334–341.
48. Shin, M.; Song, J.; Kim, M.G.; Yu, H.W.; Choe, E.K.; Chai, Y.J. Thyro-GenAI: A Chatbot Using Retrieval-Augmented Generative Models for Personalized Thyroid Disease Management. *Journal of Clinical Medicine* **2025**, *14*, 2450.
49. Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Garcia Valencia, O.A.; Cheungpasitporn, W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. *Medicina* **2024**, *60*, 445.
50. Ge, J.; Sun, S.; Owens, J.; Galvez, V.; Gologorskaya, O.; Lai, J.C.; Pletcher, M.J.; Lai, K. Development of a liver disease–specific large language model chat interface using retrieval-augmented generation. *Hepatology* **2024**, *80*, 1158–1168.
51. Long, C.; Subburam, D.; Lowe, K.; Santos, A.; Zhang, J.; Hwang, S.; Saduka, N.; Horev, Y.; Su, T.; Cote, D.; et al. ChatENT: Augmented Large Language Model for Expert Knowledge Retrieval in Otolaryngology-Head and Neck Surgery. medRxiv 2023 **2023**.
52. Yu, H.; Guo, P.; Sano, A. Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation. In Proceedings of the Proceedings of the 3rd Machine Learning for Health Symposium; Hegselmann, S.; Parziale, A.; Shanmugam, D.; Tang, S.; Asiedu, M.N.; Chang, S.; Hartvigsen, T.; Singh, H., Eds. PMLR, 10 Dec 2023, Vol. 225, *Proceedings of Machine Learning Research*, pp. 650–663.
53. Chen, R.; Zhang, S.; Zheng, Y.; Yu, Q.; Wang, C. Enhancing treatment decision-making for low back pain: a novel framework integrating large language models with retrieval-augmented generation technology. *Frontiers in Medicine* **2025**, *12*, 1599241.
54. Rani, M.; Mishra, B.K.; Thakker, D.; Khan, M.N. To Enhance Graph-Based Retrieval-Augmented Generation (RAG) with Robust Retrieval Techniques. In Proceedings of the 2024 18th International Conference on Open Source Systems and Technologies (ICOSST). IEEE, 2024, pp. 1–6.
55. Wu, J.; Zhu, J.; Qi, Y.; Chen, J.; Xu, M.; Menolascina, F.; Grau, V. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187* **2024**.
56. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R.O.; Larson, J. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* **2024**.
57. Sophaken, C.; Vongpanich, K.; Intaphan, W.; Utasri, T.; Deepho, C.; Takhom, A. Leveraging Graph-RAG for Enhanced Diagnostic and Treatment Strategies in Dentistry. In Proceedings of the 2024 8th International Conference on Information Technology (InCIT). IEEE, 2024, pp. 606–611.
58. Shi, W.; Zhuang, Y.; Zhu, Y.; Iwinski, H.; Wattenbarger, M.; Wang, M.D. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In Proceedings of the Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2023, pp. 1–10.
59. Su, C.; Wen, J.; Kang, J.; Wang, Y.; Su, Y.; Pan, H.; Zhong, Z.; Hossain, M.S. Hybrid RAG-empowered multi-modal LLM for secure data management in Internet of Medical Things: A diffusion-based contract approach. *IEEE Internet of Things Journal* **2024**.
60. Jabarulla, M.Y.; Oeltze-Jafra, S.; Beerbaum, P.; Uden, T. MedBlock-Bot: A Blockchain-Enabled RAG System for Providing Feedback to Large Language Models Accessing Pediatric Clinical Guidelines. In Proceedings of the 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2025, pp. 845–850.
61. Tayebi Arasteh, S.; Lotfinia, M.; Bressem, K.; Siepmann, R.; Adams, L.; Ferber, D.; Kuhl, C.; Kather, J.N.; Nebelung, S.; Truhn, D. RadioRAG: Online Retrieval-augmented Generation for Radiology Question Answering. *Radiology: Artificial Intelligence* **2025**, p. e240476.
62. Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.t.; Koh, P.W.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251* **2023**.
63. Jain, S.; Agrawal, A.; Saporta, A.; Truong, S.Q.; Duong, D.N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M.P.; Ng, A.Y.; et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463* **2021**.

64. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2019, Vol. 33, pp. 590–597.
65. Jin, D.; Pan, E.; Oufattole, N.; Weng, W.H.; Fang, H.; Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **2021**, *11*, 6421.
66. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146* **2019**.
67. He, X.; Chen, S.; Ju, Z.; Dong, X.; Fang, H.; Wang, S.; Yang, Y.; Zeng, J.; Zhang, R.; Zhang, R.; et al. MedDialog: Two Large-scale Medical Dialogue Datasets, 2020, [arXiv:cs.LG/2004.03329].

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.